HQSYN16 - Task #3935

Task # 3680 (New): RA4a - Automatic error prediction

Task # 3698 (Closed): Experiment with one-class clasification for join cost enhancements

Add classifier scripts to SVN

31.05.2016 16:02 - Tihelka Dan

Status:	Closed	**Start date:**	31.05.2016
Priority:	Normal	**Due date:**	05.06.2016
Assignee:	Matoušek Jindřich	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	RA1: Analysis of artifacts in synthetic speech		

Description

Please add OCC classifier scripts (anomaly_train.py, anomaly_eval.py, and other related stuff) to SVN, to start their modifications:

- output data to structured format (JSON)
- avoid using pickled data as input (it is hard to tune it with different SciKit version)

History

#1 - 10.06.2016 10:44 - Tihelka Dan

- Status changed from New to Resolved

There were several modification of the scripts carried out. Here I describe them with a few warnings:

JSON output format
All the data (except the logs) are now stored to JSON files instead of text files (in case of *-r/--cv-report* output) or pickled classifier pipe (in case of *odet* argument). There is nothing to pay special attention to, except the fact that old trained classifiers cannot be read by the new sources anymore

no pickled OCC objects
one of the most significant changes is that the (initialized) classifier, scaler and grid-search parameters are not read from a pickled object, but a python scripts from which these objects are created are passed as the 2nd and 3rd commandline argument (1st is the data provider):

```
./anomaly_train.py data_getter.py one_class_svm.py std_scaler.py one_class_svm.trained.json ...
./anomaly_eval.py  data_getter.py one_class_svm.trained.json ...
```

There are backward compatibility, though, scripts occ_pickled.py and scaler_pickled.py allow to read the old pickled data. To use them, you must call the training as:

```
./anomaly_train.py data_getter.py scaler_pickled.py occ_pickled.py one_class_svm.trained.json -S std-scaler.p
-I one-class-svm_init.p -g one-class-svm_grid.p ...
```

no pickled data objects
the second significant change is that the data are also read through Python's module, instead of the pickled numpy arrays. The module is set as the 1st command line parameter to both ./anomaly_train.py and ./anomaly_eval.p scripts, and it must provide object implementing the interface defined in data_source.py module.

There is backward compatibility script data_pickled.py, which allows to read the old pickled data. To use them, you must call the training as:

```
./anomaly_train.py data_pickled.py ... -X0 train.p -X1 train.anomaly.p -v train.cvsplit.p
./anomaly_eval.py  data_pickled.py ... -x0 eval.p -x1 eval.anomaly.p
```

Note that **the evaluation data use *-x0/-x1* command line switches**, instead of (capital) *-X0/-X1* switches which were originally used by ./anomaly_eval.p. This is the important think to keep in mind, **otherwise the evaluation will use wrong data** (those used for training)!

Each script (either providing data, scaler or clasifier) can define its own options to configure the particular module.

#2 - 13.06.2016 17:46 - Matoušek Jindřich

- Parent task changed from #3811 to #3698

#3 - 23.06.2016 16:45 - Tihelka Dan

- Status changed from Resolved to Closed